Journal of Biomedical Informatics **34,** 37–50 (2001)
doi:10.1006/jbin.2001.1001, available online at http://www.idealibrary.com on IDEAL®

Data in the Deduplication of Immunization Registry Patient Records

Perry L. Miller,[1] Sandra J. Frawley, and Frederick G. Sayward

Center for Medical Informatics, Yale University School of Medicine, New Haven, Connecticut 06520-8009

Received November 11, 2000; published online March 13, 2001

Duplicate patient records pose a major problem for many immunization registries, as well as for many electronic patient record systems. This paper reports two complementary studies exploring the deduplication of immunization registry records. One study explores the utility of different demographic data elements, singly and in combination, to assist in the deduplication process. The second study explores how clinical patient data (vaccination history data) might assist in this process. To assess the utility of demographic data elements, data were used from three registries after duplicates had been identified. A computer program, IMM/Scan, was written to count the number of true-positive (TP) matches and false-positive (FP) matches found when using different Boolean combinations of demographic data elements. In this study, a strategy of "ORing high value ANDed pairs of data elements" appeared to be most powerful. To assess the utility of vaccination history data, record pairs were drawn from 440,000 patient records. Two metrics on patient history were tested: (1) the number of identical doses shared by two records, and (2) the number of "extra" doses in the combined history of two records. In this study, sample findings include: (1) for pairs of nonduplicate records, 93% had no identical doses and 90.6% had "extra" doses, and (2) for pairs of duplicate records, 83.8% had one or more identical doses and 82% contained no "extra" doses. These studies demonstrate potentially useful approaches to using demographic data and patient history data to assist the automated deduplication of immunization patient records.

[1]To whom correspondence and reprint requests should be addressed at Center for Medical Informatics, Yale University School of Medicine, P.O. Box 208009, New Haven, CT 06520-8009. Fax: 203-764-6717. E-mail: perry.miller@yale.edu.

1. INTRODUCTION

There is currently a major national initiative to build childhood immunization registries for many different states, countries, cities, school systems, and health care organizations [1]. Particularly when records from multiple sources are pooled, for example in a statewide immunization registry, there are typically many duplicate patient records. Such a registry may have hundreds of thousands or millions of patient records, the majority of which are duplicates. As a child reaches 4 to 6 years of age, a record may include 18 or more vaccine doses involving 6 or more vaccine series.

This paper presents the results of a cooperative research project supported by the Centers for Disease Control and Prevention that involves two complementary studies focusing on the deduplication of immunization registry records. The first study explores the utility of different demographic data elements, singly and in combination, to assist in the deduplication process. The goal is to explore these utilities systematically and to assess the efficiency tradeoffs involved.

The second study explores how clinical patient data (vaccination history data) might assist in the deduplication process. Most, if not all, automated deduplication of patient records has focused on using demographic data. The authors know of no other published work describing the use of

1532-0464/01 $35.00

vaccination history data to assist in patient record deduplication. This study provides an example, illustrated using immunization data, of how the clinical content of a patient record might be used to assist in the deduplication process.

2. BACKGROUND

Duplicate records pose a potential problem for any transaction database containing demographic data. The problem is amplified by the absence of an accepted unique patient identifier. The difficulty of deduplicating records using demographic data reflects many factors, for example: (1) there are many common names; (2) a name may be spelled or expressed several ways; (3) a person's name may change; (4) addresses frequently change; (5) some registries do not record social security number for privacy reasons; (6) many children do not have a social security number; (7) fields may be blank; and (8) diverse data entry errors may occur. The problem of duplicate records is particularly important as health care systems integrate patient data from many institutions to build an enterprise master patient index (EMPI).

The presence of duplicate patient records in an immunization registry has been identified as an important problem [2]. As a registry matures, it may receive data from an increasing number of sources. The amount of missing, duplicate, and incorrect data may also increase. Problems with data quality have proven to be greater than originally anticipated. As a result, immunization registries have been forced to make data quality assurance a high priority and to allocate substantial resources to the task [3–5].

A variety of software tools and approaches have been developed to assist in the deduplication process [2, 6–8]. These apply deterministic or probabilistic matching strategies to demographic data elements, such as date of birth, last name, first name, social security number, address, parent name, and guardian name, to attempt to identify likely duplicate records.

3. EXPLORING THE UTILITY OF DEMOGRAPHIC DATA ELEMENTS

The first study explores the relative utility of different demographic data elements, individually and in different Boolean combinations, to assist in the deduplication process, using data from three immunization registries.

3.1. Three Test Immunization Data Sets

We obtained immunization data sets from three registries: a small western county (SWC) with 9728 patient records, a large western county (LWC) with 193,323 patient records, and a large eastern city (LEC) with 169,410 patient records. Both the SWC and LWC data sets allow multiple records for the same individual, so that we had a record of the previous deduplication efforts that had been carried out by these two registries. In the LWC data set, 2823 individuals have 6 or more records each. In the much smaller SWC data set, 18 individuals have 6 or more records each. In a few instances, individuals had more than 20 records. The LEC registry merges records so that a single record is maintained for each individual.

Thus each data set had already been subjected to the internal deduplication processes of the three registries. We then subjected all three data sets to further deduplication using the commercial AutoMatch 4.2 software [8] (from MatchWare Technologies, Inc.) and using MS Access database queries. AutoMatch runs and Access queries were designed for each data set to identify probable duplicates which had been missed by the registries' internal processes. All of the AutoMatch and Access results were reviewed manually and a decision was made as to which records should be identified as duplicates.

In this fashion, we created three deduplicated "test" data sets, which contained all the original records in each dataset together with an indication of which records had been identified as duplicates. Deduplication of records in a large immunization registry is always a process of incrementally approaching an idealized goal. It is impossible to determine with absolute certainty all the records that are duplicates. Even manual examination of each original patient record in the various providers' clinics could still leave considerable uncertainty. Our test data sets are therefore not expected to be fully deduplicated in an absolute sense. They represent a "best result" of significant work by the registry staff and by our research project staff using state-of-the-art software. The results described below must be interpreted in this context.

3.2. The IMM/Scan Computer Program

The goal of this study is to help understand how core demographic data elements contribute, both singly and in

combination, to the identification of the duplicate records in the test versions of three data sets. To allow us to explore this question in a comprehensive and flexible fashion, we built a computer program, IMM/Scan, to assist in the process. IMM/Scan takes as its input (1) a specified test data set, (2) a file outlining the data fields in that data set, and (3) a "predicate" indicating which data elements to use in searching for duplicates, and how those elements are to be combined. Example predicates include:

1. "date of birth" and "last name,"
2. "last name" and ("first name" or "middle name").

For a given record (R), the first predicate will identify as potential matches all records with the identical value for date of birth and last name as R. The second predicate will identify all records with the identical value for last name as R, and the identical value as R for first name and/or for middle name. (If either or both values for a field being compared are blank, they are not considered identical.)

3.3. Operation of IMM/Scan

IMM/Scan first reads in the test data set and performs certain simple data cleaning. This includes removing leading and trailing blanks, and for certain fields (such as middle name) removing periods and other punctuation. In addition, all letters are converted to upper case.

IMM/Scan then applies the specified predicate. In the process, IMM/Scan computes a variety of measures. The unit of analysis used in this processing is the record. For *each patient record* (R) in the data set, IMM/Scan computes the following measures.

1. The number of "real matches" for R. These are all the other records in the database that have been identified (by the registry staff and by our research staff) as being a duplicate record for R.

2. The number of "true positives" (TPs) matched to R by the predicate. These are all the real matches (as defined above) which are identified as possible matches for R by the predicate.

3. The number of "false positives" (FPs) matched to R by the predicate. These are all the records which are identified as possible matches for R by the predicate, which are not real matches.

The measures described above are determined for *each record*. IMM/Scan also computes the following *average* measures.

1. The average number of "real matches" (ARM) for each duplicate record in the data set.
2. The average number of "true positives" (ATP) for each duplicate record in the data set.
3. The average number of "false positives" for all records in the data set.
4. The "true positive percentage" (TP%) which is defined as follows: (ATP/ARM) $\times$ 100.

These measures were computed for all core demographic data elements in each data set, individually and in a variety of combinations.

3.4. Results

Tables 1, 2, and 3 show the results of IMM/Scan's analysis for the three data sets: SWC, LWC, and LEC, respectively. All the tables have the same basic format. Taking Table 1 as an example, the first four lines contain basic information about the data set as a whole.

1. The number of duplicate records indicates how many records have one or more duplicates in the data set.
2. The number of "singleton" records indicates how many records have no duplicate records in the data set.
3. The "average real matches (per dup)" (ARM) is defined in the preceding section, and is used as the basis for calculating the TP% in the remainder of the table.

Next, Table 1 shows a list of the core data elements in the data set. Next to each is shown the average TPs, the TP%, and the average FPs for that data element used alone as a single predicate, as well as the frequency of that data element in the data set (the percentage of records that have a non-blank value for that data element).

Following the individual data elements, Table 1 shows a list of predicates that involve pairs of data elements ANDed together, and their results. (A logical AND is represented by an ampersand, "&".) Next are several more complex predicates involving several data elements ANDed together.

Finally, the table shows a set of predicates that are Boolean expressions involving more complex combinations of several data elements, linked by both ANDs and ORs. (A logical OR is represented by a vertical bar, "|".)

3.5. Discussion

There are two underlying goals in performing this type of record duplication. These goals involve an implicit tradeoff.

TABLE 1
Results for the Small Western County (SWC) Data Set

SWC data set		Avg TPs (per Dup)	TP %	Avg FPs (per record)	Data element frequency (%)
Duplicate records:	4,742				
"Singleton" records:	4,984				
Total records:	9,726				
Average real matches (per Dup):	1.69				
Last name (LN)		1.59	94.2	8.06	100
First name (FN)		1.48	87.7	18.9	99.9
Middle name (MidN)		0.54	31.8	252	61
DOB		1.59	94.0	1.60	99.0
Sex		1.51	88.9	4,515	96
Mother's maiden name (MMN)		0.10	6.1	0.13	12
Race		0.47	27.6	4,045	67
SSN		0.16	9.3	0.03	13
Address (Addr)		0.84	49.6	4.04	96
LN & FN		1.41	83.1	0.02	
LN & DOB		1.49	88.2	0.07	
LN & MidN		0.52	30.6	0.28	
LN & Sex		1.42	83.9	3.70	
LN & MMN		0.10	5.9	0.02	
LN & Race		0.45	26.6	3.41	
LN & SSN		0.15	8.9	0.002	
LN & Addr		0.79	46.7	0.70	
FN & DOB		1.41	83.0	0.02	
FN & MidN		0.50	29.7	0.63	
FN & Sex		1.33	78.6	17.1	
FN & MMN		0.10	6.0	0.0004	
FN & Race		0.45	26.8	9.67	
FN & SSN		0.15	8.6	0	
FN & Addr		0.74	43.9	0.02	
DOB & MidN		0.52	31.0	0.04	
DOB & Sex		1.45	85.7	0.78	
DOB & MMN		0.10	6.0	0.01	
DOB & Race		0.46	27.0	0.50	
DOB & SSN		0.16	9.3	0.0002	
DOB & Addr		0.80	47.2	0.05	
Address & MidN		0.31	18.5	0.12	
Address & Sex		0.76	45.0	1.89	
Address & MMN		0.05	2.7	0.02	
Address & Race		0.28	16.7	1.44	
Address & SSN		0.08	4.5	0.001	
LN & DOB & Addr		0.74	44.0	0.03	
LN & DOB & FN		1.33	78.4	0.002	
LN & DOB & FN & Addr		0.67	39.5	0	
LN & DOB & FN & Addr & Race		0.27	15.9	0	
LN & DOB & FN & Addr & MMN		0.05	2.7	0	
LN & DOB & FN & Addr & Sex		0.62	36.9	0	
LN & DOB & FN & Addr & SSN		0.07	4.1	0	
LN & DOB & (FN | Addr | MMN | MidN | SSN)		1.44	85.2	0.041	
LN & DOB & (FN | Addr)		1.42	84.1	0.037	
LN & DOB & (FN | MMN)		1.34	79.0	0.007	
LN & DOB & (Addr | MMN)		0.82	48.7	0.037	
LN & (DOB | FN | SSN)		1.58	93.4	0.09	
LN & (DOB | FN | SSN | Addr | MidN | MMN)		1.59	93.8	0.98	
FN & (DOB | LN | SSN)		1.48	87.7	0.04	
FN & (DOB | LN | SSN | Addr | MidN | MMN)		1.48	87.7	0.67	
DOB & (LN | FN | SSN)		1.57	93.0	0.08	
DOB & (LN | FN | SSN | Addr | MidN | MMN)		1.59	93.9	0.12	
LN | DOB		1.69	99.98	9.6	
LN & DOB | LN & FN | LN & SSN | DOB & FN | DOB & SSN		1.65	97.6	0.11	

TABLE 2
Results for the Large Western County (LWC) Data Set

LWC data set		Avg TPs (per Dup)	TP %	Avg FPs (per record)	Data element frequency (%)
Duplicate records:	89,070				
"Singleton" records:	104,253				
Total records:	193,323				
Average real matches (per Dup):	4.96				
Last name (LN)		4.89	98.6	117	100
First name (FN)		4.94	99.6	318	100
Middle name (MidN)		1.95	39.3	887	37
DOB		4.95	99.7	27	100
Sex		4.96	99.9	~96,000	100
Race		4.93	99.5	~104,000	100
SSN		2.43	49.0	0.00005	31
Address (Addr)		1.84	37.1	30	93
LN & FN		4.88	98.3	0.33	
LN & DOB		4.88	98.3	0.10	
LN & MidN		1.94	39.1	1.09	
LN & Sex		4.89	98.5	59	
LN & Race		4.88	98.4	72	
LN & SSN		2.43	48.9	0.00005	
LN & Addr		1.81	36.5	1.04	
FN & DOB		4.92	99.2	0.09	
FN & MidN		1.95	39.3	2.71	
FN & Sex		4.93	99.4	308	
FN & Race		4.92	99.1	211	
FN & SSN		2.43	48.9	0.00	
FN & Addr		1.84	37.0	0.09	
DOB & MidN		1.95	39.3	0.14	
DOB & Sex		4.94	99.6	14	
DOB & Race		4.92	99.2	15	
DOB & SSN		2.43	48.9	0.00005	
DOB & Addr		1.84	37.0	0.06	
Address & MidN		0.71	14.2	0.27	
Address & Sex		1.84	37.1	16	
Address & Race		1.83	36.9	19	
Address & SSN		0.89	18.0	0.00003	
LN & DOB & Addr		1.81	36.4	0.04	
LN & DOB & FN		4.86	98.0	0.0006	
LN & DOB & FN & Addr		1.80	36.3	0.0002	
LN & DOB & FN & Addr & Race		1.80	36.3	0.0001	
LN & DOB & FN & Addr & Sex		1.80	36.3	0.0002	
LN & DOB & FN & Addr & SSN		0.89	17.9	0.00001	
LN & DOB & FN & Addr & Sex & Race & SSN		0.37	7.4	—	
LN & DOB & (FN | Addr | MidN | SSN)		4.87	98.1	0.04	
LN & DOB & (FN | Addr)		4.87	98.1	0.04	
LN & DOB & (FN | SSN)		4.86	98.0	0.0006	
LN & DOB & (Addr | SSN)		3.37	67.9	0.04	
LN & (DOB | FN | SSN)		4.89	98.6	0.44	
LN & (DOB | FN | SSN | Addr | MidN)		4.89	98.6	2.36	
FN & (DOB | LN | SSN)		4.94	99.6	0.42	
FN & (DOB | LN | SSN | Addr | MidN)		4.94	99.6	3.20	
DOB & (LN | FN | SSN)		4.95	99.7	0.19	
DOB & (LN | FN | SSN | Addr | MidN)		4.95	99.7	0.34	
LN | DOB		4.96	99.985	144	
LN & DOB | LN & FN | LN & SSN | DOB & FN | DOB & SSN		4.96	99.94	0.53	

TABLE 3
Results for the Large Eastern City (LEC) Data Set

LEC data set		Avg TPs (per Dup)	TP %	Avg FPs (per record)	Data element frequency (%)
Duplicate records:	2,857				
"Singleton" records:	166,553				
Total records:	169,410				
Average real matches (per Dup):	1.028				
Last name (LN)		0.58	56.1	222	100
First name (FN)		0.93	90.1	261	99.98
Middle name (MidN)		0.07	7.2	395	73
DOB		0.60	58.2	66	100
Sex		0.96	93.0	~83,000	99.6
Birth Hospital (Brth Hosp)		0.03	3.0	4,883	72
Birth Certificate ID (BCID)		—	0.0	0.006	14
Sequence No. of Birth (SeqNo)		0.03	2.4	~44,000	71
Guardian last name (GLN)		0.43	41.8	157	85
Guardian first name (GFN)		0.39	37.9	312	85
SSN		0.04	4.0	0.003	31
Address (Addr)		0.32	30.8	1.61	99.1
LN & FN		0.49	47.7	0.35	
LN & DOB		0.16	15.8	0.12	
LN & MidN		0.03	2.7	0.54	
LN & Sex		0.54	52.5	110	
LN & BrthHosp		0.02	1.6	8.23	
LN & SeqNo		0.02	1.8	89	
LN & GLN		0.32	30.6	34	
LN & GFN		0.30	29.2	0.72	
LN & SSN		0.04	3.8	0.001	
LN & Addr		0.20	19.3	0.31	
FN & DOB		0.53	51.6	0.12	
FN & MidN		0.07	6.8	4.96	
FN & Sex		0.86	84.0	251	
FN & Brthhosp		0.03	2.9	9.80	
FN & SeqNo		0.02	1.9	136	
FN & GLN		0.38	37.1	0.22	
FN & GFN		0.35	34.1	0.93	
FN & SSN		0.03	3.3	0.0008	
FN & Addr		0.28	27.0	0.01	
DOB & MidN		0.05	4.9	0.19	
DOB & Sex		0.56	54.1	33	
DOB & BrthHosp		0.02	1.9	2.49	
DOB & SeqNo		0.01	1.3	39	
DOB & GLN		0.19	18.0	0.09	
DOB & GFN		0.15	14.3	0.16	
DOB & SSN		0.02	1.5	0.0006	
DOB & Addr		0.17	16.1	0.04	
Address & MidN		0.02	1.8	0.013	
Address & Sex		0.31	29.7	0.81	
Address & BrthHosp		0.02	1.5	0.21	
Address & SeqNo		0.01	1.4	0.49	
Address & GLN		0.18	17.9	0.44	
Address & GFN		0.18	17.1	0.41	
Address & SSN		0.01	1.4	0.0005	
LN & DOB & Addr		0.05	4.7	0.03	
LN & DOB & FN		0.10	10.0	0.0004	
LN & DOB & FN & Addr		0.02	2.2	0.0002	
LN & DOB & FN & Addr & BrthHosp		0.0007	0.1	0.0002	
LN & DOB & FN & Addr & GFN		0.01	1.0	0.0002	

TABLE 3—*Continued*

LEC data set	Avg TPs (per Dup)	TP %	Avg FPs (per record)	Data element frequency (%)
LN & DOB & FN & Addr & GLN	0.01	1.0	0.0002	
LN & DOB & FN & Addr & Sex	0.02	1.9	0.0002	
LN & DOB & FN & Addr & SSN	0.001	0.1	—	
LN & DOB & (FN | MidN | Addr | GFN | GLN | BrthHosp)	0.15	14.8	0.047	
LN & DOB & (FN | Addr)	0.13	12.5	0.026	
LN & DOB & (FN | GFN)	0.14	13.5	0.023	
LN & DOB & (FN | GLN)	0.15	14.2	0.037	
LN & DOB & (Addr | GLN)	0.09	8.9	0.042	
LN & (DOB | FN | SSN)	0.56	54.1	0.46	
LN & (DOB | FN | SSN | Addr | MidN | GFN | GLN)	0.57	55.3	36	
FN & (DOB | LN | SSN)	0.92	89.7	0.46	
FN & (DOB | LN | SSN | Addr | MidN | GFN | GLN)	0.92	89.8	6.48	
DOB & (LN | FN | SSN)	0.59	57.5	0.23	
DOB & (LN | FN | SSN | Addr | MidN | GFN | GLN)	0.59	57.8	0.61	
LN | DOB	1.01	98.6	287	
LN & DOB | LN & FN | LN & SSN | DOB & FN | DOB & SSN	0.98	95.6	0.58	
LN & DOB | LN & FN | LN & SSN | DOB & FN | DOB & SSN | FN & SSN	0.98	95.6	0.58	

1. The primary goal is to achieve as high a true-positive rate as possible.
2. The second goal is to reduce the false-positive rate to be as low as possible, while degrading (reducing) the true-positive rate as little as possible.

It is important to keep these two conflicting goals in mind when interpreting the results.

The small western county (SWC) data set. We will discuss the SWC results (Table 1) first in some detail, and then the other results more briefly. In the fourth data column across from each single data element name is the frequency of that data element in the data set. Notice that three data elements, last name (LN), first name (FN), and date of birth (DOB), are almost always (99+%) present, and two additional data elements, Sex and Address (Addr), are each 96% present. The other data elements are frequently absent, which affects the later results in the table.

Looking first at the results for the single data elements, LN and DOB have the highest true-positive percentage (TP%), 94.2 and 94.0%, but have a quite high level of false positives (FPs), 8.06 and 1.60. As one would expect, Sex has quite a high TP% (88.9%), but a tremendously high level of FPs (4515). Looking next at the ANDed pairs of data elements, notice that when elements that by themselves have a high TP% and a high frequency are ANDed together (e.g., "LN & FN"), the results maintain quite a high TP% (although there is some degradation) and the level of FPs tends to be much lower than with the individual data elements. Looking next at the predicates where three or more data elements are ANDed together, one can see that the TP% now degrades significantly, even though the level of FPs drops to very low levels.

Looking finally at the predicates with ORs in addition to ANDs, one can see that the use of ORs has the potential to keep TP% high, while allowing FPs to be kept quite low. At one extreme, if one simply ORs together two data elements each of which have individual high TP% (see "LN | DOB"), one can achieve a very high TP% (99.98%) but at the expense of a quite high level of FPs (9.6).

Of all the strategies for combining data elements explored, the most powerful was a strategy that might best be called "ORing high-value ANDed pairs." An example of this is seen in the last line of the table. This predicate can be expressed as follows: "(LN & DOB) | (LN & FN) | (LN & SSN) | (DOB & FN) | (DOB & SSN)." As seen in Table 1, this predicate has a significantly higher TP% (97.6%) than any other predicate listed (with the exception of "LN | DOB") and achieves this with only a very modest level of FPs (0.11).

The large western county (LWC) data set. The LWC data set (Table 2) is much larger than the SWC data set discussed above, and has a high level of duplicate records. Because the database is so much larger, the number of false positives is often much larger as well. Nevertheless, the same overall pattern of results described for the SWC data

set still holds for this data set. The predicate "LN | DOB" again yields a very high TP% (99.985%) but with a very high level of FPs (144). Here again the most powerful strategy seems to be that of "ORing high-value ANDed pairs." As seen on the last line of Table 2, this predicate yields a TP% of 99.94% with a level of FPs of 0.53.

The large eastern city (LEC) data set. The LEC data set is interesting from several perspectives. First, the level of duplicate records is much lower than in the other two data sets (1.7% in the LEC data set compared to almost 50% in the LWC and SWC data sets). Also, the average number of real matches per duplicate record is also low (1.028 compared to 1.69 and 4.96 for the other data sets). This reflects the fact that the LEC registry had removed from the data set any duplicates which they had previously identified. As a result, this data set was inherently much cleaner than the other two.

Another interesting observation is that the data elements LN and DOB, when applied singly, do not have as high TP% (56.1 and 58.2%, respectively) as in the other data sets. This presumably reflects the fact that the LEC registry has already used these two data items extensively as a basis for removing duplicates from the data set. In contrast, FN has a significantly higher TP% (90.1%), presumably reflecting the fact that that element has not figured as prominently in the registry's own searching for duplicates as LN and DOB.

Although the numbers are considerably lower for this data set, the overall pattern is still quite similar to that of the other two data sets. Here again the predicate "LN | DOB" has the highest TP% (98.6%), but as expected has a high level of FPs (287). The strategy of "ORing high-value ANDed pairs" (see the last two lines of Table 3) again appears to be the most powerful overall, achieving a TP% of 95.6% while keeping the level of FPs to 0.58.

3.6. General Observations

One general observation that can be made based on these results is that the utility of any specific data element to assist in the deduplication process will vary from registry to registry, depending on factors such as (1) the frequency of that data element in the database, and (2) the nature of any deduplication that has already been performed. For example, if previous deduplication has heavily focused on LN and DOB, then these elements will not be as useful in finding new duplicates as in certain other registries, where for example less previous deduplication may have been performed.

When considering different strategies for combining several data elements in a Boolean expression (an expression involving ANDs and ORs), it is useful to bear in mind the two conflicting goals discussed previously of (1) trying to achieve a high level of TPs, while (2) trying to reduce the level of FPs as much as possible without significantly degrading the level of TPs. From this perspective, the use of ANDs alone has limited value. The data show that when several data elements are ANDed, the FP level can be reduced to very low levels but at a potentially major penalty in the level of TPs. Conversely, using ORs alone (e.g., "LN | DOB") can achieve an extremely high level of TPs, but this is accompanied by a quite high level of FPs.

As discussed previously, the most powerful strategy overall for all three data sets appeared to be that of "ORing high-value ANDed pairs." It is interesting to observe that when a registry uses a commercial deduplication tool such as AutoMatch, they tend to use it in an iterative fashion, performing multiple runs. Each run typically consists of a set of "blocking" ANDs to partition the database, followed by a probabilistic matching of one or more additional data elements. The probabilistic matching may include using "Soundex" algorithms on first and last names. Such a run can be thought of as a search involving ANDs, but with a probabilistic matching component added. In the present study, we used AutoMatch and its probabilistic capabilities to help identify duplicate records initially when constructing our test data sets, but did not include probabilistic matching in IMM/Scan. Probabilistic matching is clearly useful in practice (and would complement whatever Boolean strategy is used), but was not the focus of our study. When several different AutoMatch runs are performed and the results are examined, one is in a sense performing the ORing of several ANDed searches. One difference between this approach and that of IMM/Scan, however, is that IMM/Scan ORs several ANDed expressions in an integrated fashion, and presents the results as one integrated whole where each record appears only once.

These results suggest that a registry might continue its existing deduplication processes, but might periodically perform an integrated analysis (ORing high-valued ANDed pairs of data elements) to allow more global and integrated assistance in deduplication. It is important to point out that the OR operator has the potential to be very computationally intensive. By judicious programming, however, it was possible to structure IMM/Scan's analysis so that the Boolean predicates shown in Tables 1–3 could be performed quite efficiently. For example, the complex predicate "(LN & DOB) | (LN & FN) | (LN & SSN) | (DOB & FN) | (DOB & SSN)" was computed in 2–3 h for each of the two large

data sets using a dual-processor 550 MHz PC with 512 MB of RAM. It was important to structure the program so that each record was not compared against every other record, which could be prohibitive computationally. (This might take weeks or more to compute for a large data set.) In our programming approach, we used sorting and partitioning of intermediate results to avoid this problem.

A final comment concerns the use of the social security number (SSN) in deduplication. Although SSN is present in only 13–31% of the records in our test data sets, when it is present, it can have a significant impact in identifying TPs while reducing FPs. This confirms the potential value of a unique patient identifier in patient record deduplication.

4. EXPLORING THE UTILITY OF VACCINATION HISTORY DATA

The second study explores how vaccination history data might assist in the deduplication process. Although the problem of deduplicating vaccination history records has been addressed [9], the authors are aware of no other published work that has explored the utility of vaccination history data in immunization patient record deduplication.

One problem in designing an automated tool to compare vaccination histories is defining exactly how that tool should operate. When a human registry staff member compares two vaccination histories, that person uses a wide variety of knowledge about factors such as immunization schedules, temporal relationships, and likely types of data errors, to make an intelligent assessment as to whether they might involve the same patient. At the same time, making such an assessment for many records would be very time-consuming and would likely be performed inconsistently.

This paper explores how two heuristics (computational "rules of thumb") can be applied to help determine whether two immunization histories involve the same patient or not. As described later in the paper, the first heuristic counts the number of "identical doses" in the combined history of two records, and the second heuristic calculates an "extra dose penalty" when the combined history contains too many doses for the patient's age. A software tool can apply these heuristics automatically, consistently, and very rapidly.

4.1. Selecting "Matching" and "Nonmatching" Pairs of Patient Records

This study was performed using the database of a large state immunization registry, containing roughly 430,000 patient records, prior to that state's own record deduplication efforts. This database was a merged set of records provided by a variety of sources, including insurance companies, health maintenance organizations, and individual providers.

We first built a software tool designed to randomly select "matching" and "nonmatching" pairs of patient records based on demographic data. This tool allows the user to specify (1) an age range, (2) a set of demographic data items to consider, (3) the number of record pairs desired, and (4) whether matching or nonmatching records pairs are desired. Using this tool, we created a total of 6 sets of randomly selected record pairs. One matching set of pairs and one nonmatching set of pairs were identified for each of the following three age ranges: 12–23 months of age, 24–35 months of age, and 36–47 months of age. Each set contained 1000 record pairs.

Matching pairs. Each matching pair of records was selected with the requirement that each record have the identical birth date, first name, and last name. We also checked to make sure that if both records had values for middle name, mother's maiden name, or social security number, that these were also identical in both records. The goal was to assure, as much as possible, that each pair of records involved the same patient. In addition, the vaccination history of each record was required to contain at least one DTP series dose and one polio series dose.

Nonmatching pairs. Each nonmatching pair of records was selected with the requirement that each record have the identical birth date, a different first name, and a different last name. Each nonmatching pair was inspected to assure that there was nothing in the patients' names to suggest that they might be the same patient (for example, minor spelling differences in the first and last names). The goal was to assure, as much as possible, that each pair of records involved different patients who had the same birth date. In addition, the vaccination history of each record was required to contain at least one DTP series dose and one polio series dose.

4.2. Applying the History Comparison Heuristics

To carry out the analysis, we wrote a computer program that performed the steps outlined below. The first step involved inspecting the vaccination history associated with each individual record and removing "duplicate" doses. A duplicate dose pair was defined as two doses in the same vaccine series recorded as given on the same date. For example, if a patient record indicated that a DT dose and a DTaP dose were each given on 3/1/99, then one of those doses was arbitrarily selected and removed.

Heuristic 1. Once these duplicate doses had been removed from each individual record, each pair of records was inspected to count the number of "identical" DTP doses and polio doses that were contained in the combined vaccination history of both records. Two doses were considered "identical" for this purpose if they were in the same vaccine series and had the same date. Only the DTP and polio series were considered. For example, the following record pair would be considered to have two identical DTP series doses and one identical polio series dose (see italics).

Record 1: DTP *1/1/98*, 4/1/98, *7/1/98*; IPV *1/1/98*, 4/1/98
Record 2: DTaP *1/1/98*; DTP *7/1/98*; OPV *1/1/98*, 5/1/98

Heuristic 2. The second heuristic was also calculated using the DTP and polio vaccination histories of each record pair. This heuristic was designed to help assess if the combined history of both records had more doses for a series than one would expect at several specified ages, and if so, to compute an "extra dose penalty."

The top portion of Table 4 shows the values used to compute this penalty for DTP. The first column lists several ages (expressed in weeks, months, or years) and the second column indicates the expected number of DTP series doses that a child should normally have received by that age. If the combined history contains more doses than expected for a given age, the next columns indicate a dose penalty to be assigned. The higher the number of extra doses, the higher the penalty (which is expressed as a negative number).

For example, by age 13 weeks, a child is expected to have received no more than two DTP series doses. If three doses were recorded in the combined history before this age, then a penalty of −5 is assigned. If four doses were recorded before this age, then a penalty of −30 is assigned, and so on. The combined history for a record pair is analyzed in this fashion for each age specified in the table. The maximum penalty assigned for any age is then assigned to the pair. The bottom portion of Table 4 shows the values used to calculate the extra dose penalty for the polio vaccine series.

Inspection of anomalous record pairs. After examining the results of applying the two heuristics, a small number of matching record pairs were identified as "anomalous." These records appeared to be the same patient based on demographic data, but the results of applying the heuristics suggested either (1) that two different patients were involved or (2) that a data error had occurred. We examined both the demographic data and the vaccination histories of these anomalous matching pairs, and prepared a table outlining the apparent explanation for the anomalies.

4.3. Results

Identical doses. Table 5 shows the results of applying heuristic 1, counting "identical" doses in each pair of records. The results are presented first for the DTP series and then for the polio series. Finally, combined figures are shown. The four columns on the left provide the results for the 3000 matching pairs. For each observed identical dose value, we first show the percentage of the 3000 matching pairs which had that number of identical doses. The next three columns show the actual number of pairs for each age range. The rightmost four columns provide similar figures for the 3000 nonmatching record pairs. In comparing the results for matching pairs vs nonmatching pairs, there is a striking

TABLE 4

	Age	Expected doses	Extra dose penalty		
			+1 dose	+2 doses	+3 doses or more
DTP series	9 weeks	1	−5	−30	−40
	13 weeks	2	−5	−20	−40
	11 months	3	−5	−15	−40
	3 years	4	−5	−10	−40
	6 years	5	−5	−5	−40
Polio series	9 weeks	1	−5	−30	−40
	13 weeks	2	−5	−20	−40
	3 years	3	−5	−15	−40
	5 years	4	−5	−10	−40

Note. The top portion of this table shows the "extra dose penalty" assigned to a combined DTP series history if it contains more doses than expected at any specified age. For each combined history of a record pair, the penalty is calculated as described in the text. The bottom portion of the table shows the same information for polio.

TABLE 5
Identical Doses in Matching and Nonmatching Record Pairs

# Identical Doses	Matching record pairs: % (12–47 months)	12–23 months	24–35 months	36–47 months	Nonmatching record pairs: % (12–47 months)	12–23 months	24–35 months	36–47 months
DTP Series								
0	16.3	202	138	149	93.1	925	940	929
1	28.5	364	263	229	6.4	67	55	69
2	22.9	217	233	237	0.5	8	5	2
3	21.6	195	251	201	—	0	0	0
4	10.7	22	115	183	—	0	0	0
5	—	0	0	1	—	0	0	0
Total	100	1000	1000	1000	100	1000	1000	1000
Polio series								
0	17.1	204	146	162	93.7	930	947	933
1	30.9	373	282	273	6.0	64	50	65
2	26.3	256	264	269	0.4	6	3	2
3	24.7	166	299	277	—	0	0	0
4	0.9	1	9	18	—	0	0	0
5	—	0	0	1	—	0	0	0
Total	100	1000	1000	1000	100	1000	1000	1000
DTP & polio								
0	16.2	200	137	148	93.0	925	939	925
1	0.9	6	9	11	0.8	4	9	11
2	27.1	355	242	215	5.8	64	47	62
3	4.2	18	50	58	0.1	1	2	1
4	19.3	203	194	183	0.3	6	3	1
5	7.0	52	62	95	—	0	0	0
6	15.0	147	195	108	—	0	0	0
7	9.5	18	102	165	—	0	0	0
8	0.9	1	9	16	—	0	0	0
9	—	0	0	1	—	0	0	0
Total	100	1000	1000	1000	100	1000	1000	1000

difference. For example, 93% of the nonmatching pairs had no identical doses at all in these vaccine series. In contrast, 83.8% of the matching pairs had one or more identical doses.

Extra dose penalty. Table 6 shows the results of applying heuristic 2, calculating an "extra dose penalty" for each pair of records. The format of this table is similar to that of Table 5. In comparing the results for matching pairs vs nonmatching pairs, there is again a major difference. Here, 82.3% of the matching pairs had a penalty of zero. Only 9.4% of the nonmatching pairs had a penalty of zero, while 45% had a penalty of −35 to −80 when the series penalties were combined.

Combining the two heuristics. Figure 1 shows a matrix which correlates the results for identical doses against the results for extra dose penalties, using the numbers from Tables 5 and 6 for DTP and polio combined. For each cell in this matrix, Fig. 1 first shows the relevant percentage of matching pairs, followed by the percentage of nonmatching record pairs. For example, the top-left cell indicates that 9.6% of matching record pairs and 8.2% of nonmatching record pairs have zero identical doses and a zero penalty. To help better understand these numbers, we have shaded two sets of cells, one set with dark shading and the other set with light shading.

Inspection of the dark-shaded cells shows that these pairs are highly likely to be nonmatching. Here, 88.4% of the nonmatching pairs fall in these cells, while only 6.7% of the matching pairs do. We have therefore labeled these cells "likely to be different patients." Similarly, pairs in the lightly shaded cells are highly likely to be matching. Here, 78.5% of the matching pairs fall in these cells, while only 1.5% of the nonmatching cells do. We have therefore labeled these cells "likely to be the same patient."

TABLE 6
"Extra Dose" Penalty for Matching and Nonmatching Record Pairs

Penalty	Matching record pairs: % (12–47 months)	Matching record pairs: 12–23 months	Matching record pairs: 24–35 months	Matching record pairs: 36–47 months	Nonmatching record pairs: % (12–47 months)	Nonmatching record pairs: 12–23 months	Nonmatching record pairs: 24–35 months	Nonmatching record pairs: 36–47 months
DTP series								
0	85.8	888	821	864	12.5	149	105	122
−5	10.9	103	127	97	23.7	256	231	223
−10 to −15	2.4	7	40	24	22.1	267	203	193
−20 to −40	1.0	2	12	15	41.7	328	461	462
Total	100.0	1000	1000	1000	100.0	1000	1000	1000
Polio series								
0	84.3	893	800	836	12.5	171	97	108
−5	11.6	99	133	116	27.7	344	248	240
−10 to −15	3.3	6	56	37	26.0	279	249	252
−20 to −40	0.8	2	11	11	33.7	206	406	400
Total	100.0	1000	1000	1000	100.0	1000	1000	1000
DTP and polio								
0	82.3	874	784	811	9.4	125	74	83
−5 to −10	13.1	115	143	134	23.0	258	217	215
−15 to −20	1.8	5	27	22	9.6	114	87	88
−25 to −30	1.8	4	33	17	13.0	150	127	113
−35 to −80	1.0	2	13	16	45.0	353	495	501
Total	100.0	1000	1000	1000	100.0	1000	1000	1000

The top left cell (zero identical doses and zero penalty) is interesting because roughly the same percentage (8–9%) of matching and nonmatching record pairs falls in this cell. On inspection, these tend to be record pairs where each record contains only a portion of a normal vaccination history for the DTP and polio series.

Inspection of anomalous matching record pairs. To better understand why a small number of matching record pairs had high extra dose penalties, we examined the demographic information and the individual vaccination histories of 136 of these record pairs in detail. Table 7 summarizes the results. In 36 record pairs, one history used a default number for the day-of-month (e.g., all doses were recorded as given on the first of the month), while the other history was identical but had different numbers for day-of-month. Nineteen pairs had histories that were identical except for slight date variations; for example, the day of the month for each dose in one history might be one day later than in the other history. Other data errors included likely keystroke error, early dates (before the recorded date of birth), or otherwise invalid dates

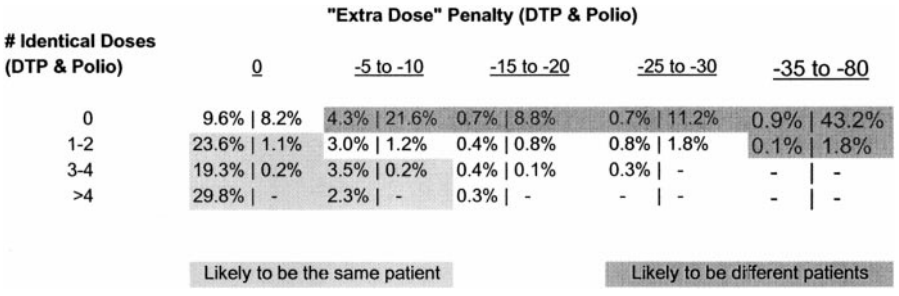

FIG. 1. A matrix correlating the results for the two heuristics, using the numbers for DTP and polio combined from Tables 5 and 6. All age groups are combined. For each cell in this matrix, the relevant percentage of matching record pairs is first displayed, followed by the percentage of nonmatching pairs.

TABLE 7
Summary of the Results of Our Manual Analysis of Anomalous Matching Pairs, as Described in the Text

Reason for anomaly	Number of record pairs
Default day-of-month used	36
Slight day-of-month variation	18
Likely keystroke error	19
Early and/or invalid dates	17
Totally different histories	42
Partially different histories	23
Other data anomalies	1
Total:	156

Note. 136 pairs were analyzed. 20 of these exhibited two problems.

for doses. For 65 pairs, the two records had totally or partially different histories. Theoretically, these pairs might involve different patients with the same first name, last name, and date of birth, although there was generally enough demographic information present to make this seem very unlikely. After inspection, our assessment was that in most if not all cases, a data entry or administrative error had probably associated incorrect history data with one of the patient records.

4.4. Discussion

The results described above in Section 4 suggest that the two heuristics have the potential to provide useful assistance in the immunization record deduplication process, particularly as an adjunct to deduplication based on demographic data. In this regard, it is useful to consider the following two scenarios.

1. If a set of high probability duplicate record matches have been identified based on demographic data, but a few percent of these matches are expected to involve different patients, the proposed approach could help identify many of these false matches.
2. If the registry has set its demographic matching filter to operate coarsely, thereby identifying a range of relatively low probability matches, most of which are likely not to be true duplicates, the proposed approach could help selectively focus on the true matches.

Limitations of the present study. The potential value of the approach needs to be assessed in the context of various limitations inherent in the present study.

1. The data used were derived from a single state registry prior to its own deduplication efforts. As a result, the numbers and percentages obtained may well differ if the approach is implemented with other registries, which may have different procedures for data entry and validation and/or might be further along in the record deduplication process.
2. For very young children, for example between 0 and 3 months of age, the approach may well not be as useful since much less history will be available for comparison. We were not able to find a sufficient number of matching record pairs for very young patients in our database to allow us to perform this analysis. For the three age groups we did study, there were not major differences between the performance of each heuristic in the different age groups. The older matching-pair groups did have somewhat higher numbers of identical doses, e.g., more pairs had three to five identical doses in each series, reflecting the probability that older children have more vaccinations recorded. This was not a factor in discriminating matching from nonmatching pairs, however, since *none* of the nonmatching pairs had more than two identical doses in a series.
3. The present study focused on the DTP and polio vaccine series. These two series were chosen because they have a relatively large number of recommended doses, compared for example to Varicella and MMR. Hib would also have been a reasonable choice. Hepatitis B would not have been as useful since dose 1 is typically given on the day of birth. The results for DTP and polio shown in Tables 5 and 6 suggest that the analysis of the two series tends to provide quite similar information about each record pair. This in turn suggests that including additional series might not yield much new information.
4. As described previously, another study restriction was the requirement that each record have at least one DTP series dose and at least one polio series dose. If one member of a record pair does not have any vaccination history recorded, then the two histories cannot be compared. In a registry, there will likely be some records with little or no history recorded.

Possible future enhancements of the approach. The computer cannot realistically be expected to emulate the behavior of a human focusing sustained attention on comparing two vaccination histories. Nevertheless, the two heuristics described in this paper have clear potential value, and a number of enhancements to the approach could be made.

The most ambiguity in Fig. 1 occurs in the top left cell (zero identical doses and zero penalty). Here, the histories of each record pair tend to have only a portion of the entire history. One enhancement that might help partially resolve this ambiguity would be to check the intervals between the doses in each series to see how many were too close together

in the combined history. If doses were too close together, this would suggest that different patients were involved.

Another enhancement would be to examine the combined histories for certain of the types of data errors described in the section on anomalies. For example, if the histories of the record pair have doses with the same month and year, but one history always indicates the first day of the month, and the other history indicates different days, then this would suggest that the patient might be the same, even though there might be no identical doses and a high extra dose penalty.

ACKNOWLEDGMENT

This research was supported in part by cooperative grant U1W/CCU114707 from the Centers for Disease Control.

REFERENCES

1. Cordero JF, Guerra FA, Saarlas KN, editors. Developing immunization registries: experience from the All Kids Count Program. Am J Prev Med 1997; 13(Suppl 1):entire issue.
2. DuBose SS, Wheeler W. Verification and validation procedures for immunization registries. Am J Prev Med 1997; 13(Suppl 1):62–65.
3. All Kids Count Program, Centers for Disease Control and Prevention. Community immunization registry manual. USDHHS, PHS, CDC, March 14, 1997.
4. Payne T, Kanvik S, Seward R, Beeman D, Salazar A, Miller Z, Immanuel V, Thompson RS. Development and validation of an immunization tracking system in a large health maintenance organization. Am J Prev Med 1993; 9:96–100.
5. Wilton R, Pennisi AJ. Evaluating the accuracy of transcribed computer-stored immunization data. Pediatrics 1994; 94:902–6.
6. Clark DE, Hahn DR. Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry. Proc Annu Symp Comput Appl Med Care 1995; 397–401.
7. Newman TB, Brown AN. Use of commercial record linkage software and vital statistics to identify patient deaths. J Am Med Inform Assoc 1997; 4:233–7.
8. Jaro MA. Probabilistic linkage of large public health data files. Stat Med 1995; 14:491–8.
9. Miller PL, Frawley SJ, Sayward FG. IMM/Scrub: A domain-specific tool for the deduplication of vaccination history records in childhood immunization registries. Comput Biomed Res 2000; 33:126–143.